Г Р А М А Т И К О Л О Г І Я

УДК 81'322

Oksana Zuban, Ph D, Doc.
Taras Shevchenko National University of Kyiv, Kyiv

MORPHEMIC AND DERIVATIONAL ANALYSIS IN THE CORPUS OF THE UKRAINIAN LANGUAGE

The morphemic and derivational analysis of the corpus of the Ukrainian language – is a convenient linguistic tool, which in an online mode helps the user to carry out the research on the study of morphemes and derivation on the basis of a great number of illustrative textual materials of the corpus of the Ukrainian language, which enables to get new knowledge about the semantic and formal structure of the Ukrainian word. Morphemic-Derivational Data Base (190 thousand lexemes are segmented) also enables: to carry out various classifying analyses of vocabulary according to the quantitative and morphic (www.mova.info).

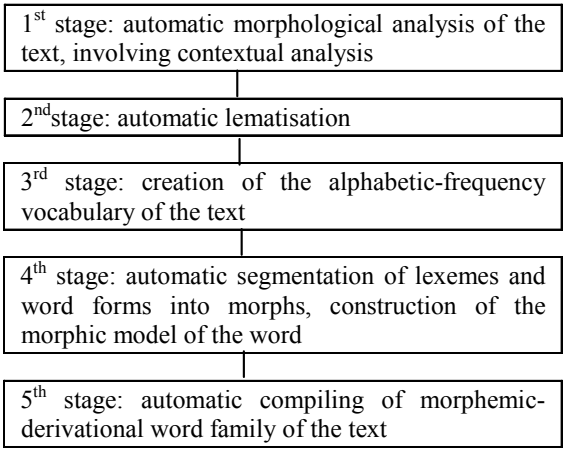
Key words: *Morphemic-Derivational Data Base, corpus of the Ukrainian language, the morphic segmentator of the Ukrainian text, Electronic dictionary of frequency, морphemic and derivational analysis, derivational word family.*

One of the tasks of the applied linguistic is getting knowledge about the structure of the language and the way it functions on different languages layers. The text as result of speech act contains a definite inventory of language elements, which are selected and combined in it in accordance with the grammar laws of the language and are regulated by the norm. Thus, it is right and timely to study the language phenomena indirectly through the text, where all the linguistic units of all the languages layers are analysed as single reality under observation. New information technologies, which are designed-in the corpus of the national languages, facilitate the selection of language phenomena and secure the process of receipt of objective/impartial data about the way language units function in speech, expand the wealth of methods and tools applied in linguistic research, improve the researcher's efficiency. Automatic segmentator of the Ukrainian text is an example of such a technology, while the accumulated corpus of the language, that counts 30 million word forms, will give the science of the Ukrainian language the invaluable

materials about the way morphemes function in different types of Ukrainian discourse.

The morphic segmentator of the Ukrainian text – is a system, on the input of which there are lexemes (or word forms) of an analysed text. They are presented in a form of an alphabetic-frequency dictionary. On its output there are the same lexemes (word forms) that are index-linked by the codes of grammatical belonging to a definite part of speech and are split into morphs – root morphs, affixal morphs with a proper index.

The stages of the segmentation of an analysed text into morphs can be presented in a form of a such sequence:



1st stage – automatic morphological analysis (AMA) – is an integral part of a linguistic guarantee of an every system dealing with automatic processing of textual information.

2nd stage – lematisation, or the output of an initial form for every textual word form. The solution of this task ensures the possibility of appealing to the dictionary of initial forms of words and paradigmatic forms, that are split into morphs.

On the **3rd stage** the alphabetic-frequency dictionary is compiled.

On the **4th stage** the segmentation of lexemes and word forms into morph takes place.

On the **5th stage** derivational word families are constructed.

The segmentation into morphs starts with the *retransmission* literal record into a simplified phonemic record according to a special procedure/ algorithm, which takes into account only the positions of sounds that are pronounced with iotacism: **я**, **ю**, **є**, **й**: makes the conversion **я → ja**, ю → **jy**. This procedure is obligatory, since it makes it possible to define the boundary of morphs, when the one-letter spelling for two sounds appear at the junction of morphs: *діяти → дій-а-ти; фантазія → фантазій-а.* All the other peculiarities of a phonemic record are not taken into account, specifically palatalized phonemes (spelling is preserved: біль), word stress, since a simple text, where there are no stress marks, is analysed, etc.

Word forms are segmented into morphs according to the conventional theoretical principles of highlighting the morphs by means of automatic confrontation with a linguistic model of a morphic structure of a word in Morphemic-Derivational Data Base (MDDB), where 190 thousand of lexemes are segmented including onomastic and toponymic lexemes. After the completion of the procedure of segmentation into morphs the word form structure of a word is memorised in symbols: R – root, P – prefix, S – suffix, F – inflection, X – postfix, I – infix. Representation of each morphic structure of a word in symbols enables to describe each morphic structure automatically with the help of a software procedure, e. g. *застудити* / P2R6S7F9 (***за*** – P2, i.e. that two initial phonemes are prefix phonemes, ***студ*** – R6 i.e. that the third, fourth, fifth and sixth phonemes are root phonemes, ***и*** – S7, the seventh phoneme is the suffix phoneme, ***ти*** – F9, i.e. the eighth and ninth phonemes are inflection phonemes). The morphic structure, which is automatically formed with the help of a software procedure, gives full linguistic information about a morph, its structural and distributional relations with other morphs and it is defined as a working unit of a dictionary of morphs.

Automatic morphic analysis enables to get information about homonymy and number of allophones of a root morph of an analysed word. An index from the list of homonymic roots is ascribed to every root of an analysed root of a language corpus, the invariant form is ascribed to the allomorph. E.g. the root ***сон*** has 3 homonymic roots: **сон$_1$** (*сонце*); **сон$_2$** (*сон*); **сон$_3$** – (*рослина*). The root **сон$_2$** has the index **$_2$**, the word *сон*, has the same index number for an allomorph **сонь** in a word *сонько.* The invariant **сон**, as well as the index**$_2$** is

ascribed to the root ***сонь*** in a word *сонько*. At the stage of a formation of a morphic base a proper index was manually ascribed to all of the derivatives of these three roots. This information is given in a Morphic-Derivational Data Base, where homonymic roots (ca. 3 thousand) and root allomorphs (ca. 2900) are marked. Proper names (*Боб* – proper name, *боб* – common noun) do not fall into the category of homonymic roots.

Automatic segmentator into morph is used linguistic studies of morphemic and derivational structure of words, particularly it is used in such studies:

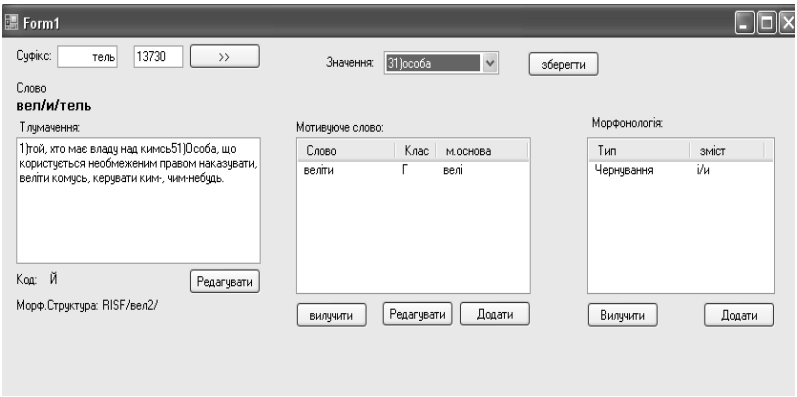
- in the process of compiling of alphabetic-frequency dictionaries that contain all types of morphs on the basis of the texts of different types;
- ascertainment and merge of allomorphic affixes into a morpheme (number of allophones of the roots is transferred automatically from the basic dictionary to the word form from the text);
- ascertainment of systemic and functional characteristics of the morphs;
- automatic construction of derivational word families (the index of homonymic roots is transferred automatically from the basic dictionary to the word form of the text);
- carrying out the derivational analysis of the word forms.

The formalised description of a morphic structure, which is represented through the linguistic model of a software procedure, enables to carry out the modeling of structural relations between forms in the two planes of an arrangement of a word as a language sign: formalization of plane of expression and plane of content. It enables to carry out the comprehensive distributional and statistical description of morphemes taking in account all their meanings and allomorphic realizations and it enables to compile the dictionary of morphemes of the Ukrainian language.

The work on the compiling of a dictionary of affix morphemes was started within the project of the Morphemic and derivational analysis in the corpus of the Ukrainian language. This work is carried out by the linguists in an automated on-line mode in order to develop the data base of word form meanings, this data base can be regarded as a base of knowledge about invariants and variants of affixes of the modern

Ukrainian language. This data base contains the list of all affixes of the Ukrainian language together with the information about their structural position, semantic structure and world building facilities (see pic.1).

The dictionary is at the stage of compiling and accumulation of knowledge about every affix in an every word of general basic dictionary of the Ukrainian language, which contains 200 thousand words. This dictionary will be used as a data base to form the users' requests connected with semantic structure of affixes, with the phenomenon of morphic juncture in derivatives, to enable automatic formation of morphemic-derivational word family of the analysed texts, which are represented in the corpus of the Ukrainian language. It is not ruled out that one will be able to use these data to carry out the semantic analysis of the text in future.



Pic. 1. Interface of the program of compiling of the dictionary of derivational meaning of affixes in operation

Automatic morphic analysis functions as the linguistic classifier in the process of compiling of an electronic derivational dictionary. The system of automatic derivational analysis is designed on the basis this linguistic classifier. The formation of an derivational word family as item of the electronic derivational dictionary is carried out on the basis of selection of all words from the analysed text having the same root. The formation of selections of words having the same root is a difficult and laborious task, thus it is necessary to formalise

the material on all the stages of its description, it enables to create software tools of the linguistic analysis.

Taking into account such principles of derivation as:

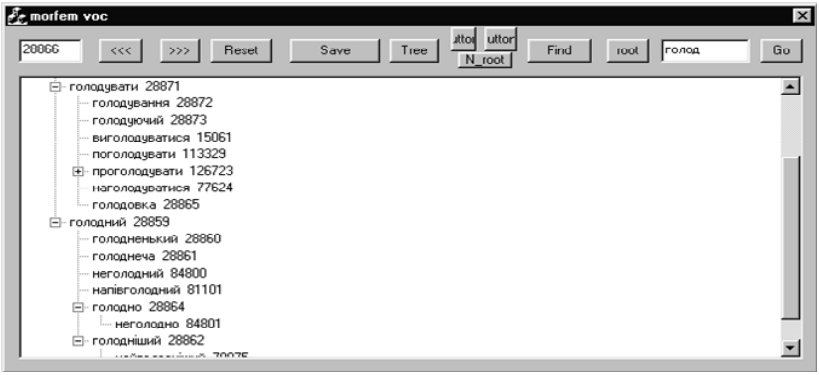
1) Morphological means of word building envisage the quantitative and affixal increase of morphic structure of the motivated base of a derivational pair;

2) Infixes are not regarded to be derivational formants and are added to the derivational suffixes and prefixes in the process of word building;

3) Compound words belong mainly to the first stage of word building,

formalising principles of description of derivational relations between motivating and motivated words were designed. It enables to build the hypothesis-model in operation of a derivational word family. In this hypothesis-model every following word building act represents words with more complex in terms of quantity affixal structures of stems. i.e., that a group of words of an each quantitative and affixal model of a word is a hypothetic stage of a derivational word family.

The automatic construction of a derivational word family is carried out by means of toolset of an electronic card (pic. 2). The word is typed in the field ***root***, that is considered to be the top of a derivational word family (i.e. the first motivating word of the words having one root). Then one presses the button ***Find*** in order to carry out the automatic grouping of the vocabulary and the classification according to the word building stages of a hypothetic derivational word family. It appears in a form of tree dependencies on the screen of the card.



Pic. 2. Interface of the derivational word family

This example demonstrates the part of the derivational word family of the words with the root -*голод*-, which is constructed in the automatic mode on the basis of the selection of the words having one root of the morphemic and derivational data base of the Ukrainian language. Each brunch of the derivational tree reflects the relations of derivational motivation between the main word, which is marked by a square with either + or-, and the words that finish the brunches of this main word. Marker "+" denotes that the word is main, i. e. that it is a starting point of the brunch, while the marker "–" denotes that this brunch is already extended. The modeling structural and motivational relations between the words of neighbouring derivational stages is carried out by means of establishing correspondences between numeric cods of words which belong to the morphemic data base: *голодувати* → *голодування, голодувати* → *поголодувати.*

The classification of the selection of words having the same root according to the derivational stages is only a linguistic hypothesis and it needs to be checked. Such tasks are set:

- to check if the words are grouped into selections of words having the same root correctly, and to check the placement of them according to the derivational stages;
- to define the derivational base and derivational formant in every derivative;
- to add information about morphonological processes, that take place in every derivational stage.

On this stage the work is carried out by a linguist, who uses his knowledge as an expert philologist and who also edits the derivational word family.

Thus, the morphemic and derivational analysis of the corpus of the Ukrainian language – is a convenient linguistic tool, which in an online mode helps the user to carry out the research on the study of morphemes and derivation on the basis of a great number of illustrative textual materials of the corpus of the Ukrainian language, which enables to get new knowledge about the semantic and formal structure of the Ukrainian word. It also enables:

- to carry out various classifying analyses of vocabulary according to the quantitative and morphic models;

- to form root dictionaries, dictionary of affixes and derivational dictionaries of different styles and discourses;
- to carry out the morphic analysis of now inlet word forms.

Стаття надійшла до редакції 12.04.14

Оксана Зубань, канд. філол. наук, доц.
КНУ імені Тараса Шевченка, Київ

Морфемний і словотвірний аналіз у Корпусі української мови

У статті описано створення і принципи роботи Морфемно-словотвірної бази даних (МСБД) української мови, яка виконує функцію модуля-аналізатора в Корпусі української мови. Реєстр словника бази об'єднує 190 тис. слів української мови і включає пакети програм, здатні виконувати різні задачі в автоматичному режимі в проекції на параметризований текст: здійснювати морфемну сегментацію словоформ; групувати лексику у спільнокореневі та спільноафіксальні вибірки; класифікувати лексику за структурно-морфними моделями; створювати алфавітно-частотні кореневі та афіксальні словники, викладені на інтернет-порталі www.mova.info.

Ключові слова: *Морфемно-словотвірна база даних, Корпус української мови, морфемний сегментатор українського тексту, електронний частотний словник, морфемний і словотвірний аналізи, словотвірне гніздо.*

Оксана Зубань, канд. филол. наук, доц.
КНУ имени Тараса Шевченко, Киев

Морфемный и словообразовательный анализ в Корпусе украинского языка

В статье описывается создание и принципы работы Морфемно-словообразовательной базы данных (МСБД) украинского языка, исполняющей функцию модуля-анализатора в Корпусе украинского языка. Реестр словаря базы представляет 190 тыс. слов украинского языка и включает пакеты программ, способные выполнять ряд задач в автоматическом режиме в проекции на параметризированный текст: производить морфемную сегментацию словоформ; группировать лексику в общекорневые и общеаффиксальные выборки; классифицировать лексику по структурно-морфным моделям; создавать алфавитно-частотные корневые и аффиксальные словари, представленные на интернет-портале www.mova.info.

Ключевые слова: *Морфемно-словообразовательная база данных, морфемный анализ, словообразовательный анализ, Корпус украинского языка, морфемный сегментатор украинского текста, словообразовательное гнездо.*